Structure from Periodic Motion

Serge Belongie and Josh Wills

Department of Computer Science and Engineering,
University of California, San Diego,
La Jolla, CA 92093
http://vision.ucsd.edu

Abstract. We show how to exploit temporal periodicity of moving objects to perform 3D reconstruction. The collection of period-separated frames serve as a surrogate for multiple rigid views of a particular pose of the moving target, thus allowing the use of standard techniques of multiview geometry. We motivate our approach using human motion capture data, for which the true 3D positions of the markers are known. We next apply our approach to image sequences of pedestrians captured with a camcorder. Applications of our proposed approach include 3D motion capture of natural and manmade periodic moving targets from monocular video sequences.

1 Introduction

Periodic motion is ubiquitous in the physical world, from the oscillations of a pendulum to the gallop of a horse. The periodicity of moving objects such as pedestrians has been widely recognized as a cue for salient object detection in the context of tracking and surveillance, see for example [1, 11]. In this paper we focus on the use of periodicity for a different and, to our knowledge, novel purpose: 3D reconstruction. The key idea is very simple. Given a monocular video sequence of a periodic moving object, any set of period-separated frames represents a collection of snapshots of a particular pose of the moving object from a variety of viewpoints. This is illustrated in Figure 1. Thus each complete period in time yields one view of each pose assumed by the moving object, and by finding correspondences in frames across neighboring periods in time, one can apply standard techniques of multiview geometry, with the caveat that in practice such periodicity is only approximate. In this paper we present this idea and apply it to the problem of estimating sparse 3D structure and dense disparity for walking humans.

The organization of this paper is as follows. We review related work in Section 2. In Section 3 we discuss our approach. Experimental results appear in Section 4, and we conclude and discuss future work in Section 5.

2 Related Work

Periodicity is a kind of symmetry, and as such, its use in recovering 3D information is related to approaches that leverage other kinds of symmetry. An early

W.J. MacLean (Ed.): SCVMA 2004, LNCS 3667, pp. 16–24, 2006.

Fig. 1. Illustration of periodic motion for a walking person. Equally spaced frames from one second of footage are shown. The pose of the person is approximately the same in the first and last frames, but the position relative to the camera is different. Thus this pair of frames can be treated approximately as a stereo pair for purposes of 3D structure estimation. Note that while the folds in the clothing change over time, their temporal periodicity makes them rich features for correspondence recovery across periods.

example of work in this vein is Kanade's method of recovering 3D shape from a single view of a skew symmetric object [9]; more recent extensions of these ideas appear in [6, 4]. The periodicity we are concerned with is temporal; in contrast, spatial periodicity (together with homoegeneity and isotropy) has been exploited in several shape-from-texture approaches, e.g. [5, 14], in which the periodicity pertains to texture elements on the surface of a curved object. While the periodicity of walking humans and animals has indeed been used for other purposes, e.g. pedestrian detection [1], to our knowledge the present work is the first to exploit it for 3D reconstruction.

3 Our Approach

In this section we describe our approach to estimating structure from periodic motion (SFPM). In illustrating the idea, we make use of motion capture (or *mocap*) data from [16]. We provide experimental results on regular video sequences in the following section.

3.1 Estimating the Period

In the present work we specify the period of the moving target manually. A number of approaches exist for estimating the period of a walking figure, e.g. [1]. As our focus is on the reconstruction problem, we have not investigated the use of these algorithms, though we do address the issue of error in the period estimation step in Section 4.

3.2 Multiview Geometry Across Periods

The most elementary configuration for periodic structure from motion is the case of two views separated in time by one period. As is well known from [2, 7], the 3D structure of a rigid object can be estimated up to a projective transformation from two uncalibrated views. The periodic motion counterpart to this is illustrated in Figure 2(a,b), which depicts two 2D views of mocap data spaced apart one period T_o in time.

In this case, the camera is stationary and the walking figure has translated and rotated relative to the camera over the course of the period. These two views correspond approximately to a stereo pair of a particular pose of the walking figure. The reconstruction obtained from these two views is shown in Figure 2(c). Since we are using uncalibrated cameras, the reconstruction is arbitrary up to a 3D homography; our display shows the reconstruction using a least-squares homography estimated using the ground truth marker positions. Alternatively, if three or more views are available, one can employ autocalibration techniques such as [13]. Partial calibration information can also be obtained from knowledge about the scene (see e.g. [8] Ch. 18) or from known properties of the moving target, e.g. that it is a human of a certain aspect ratio.

As is the case in standard structure from motion (SFM), the underlying geometry is only part of the problem: one must solve for correspondences between views before estimating the structure.

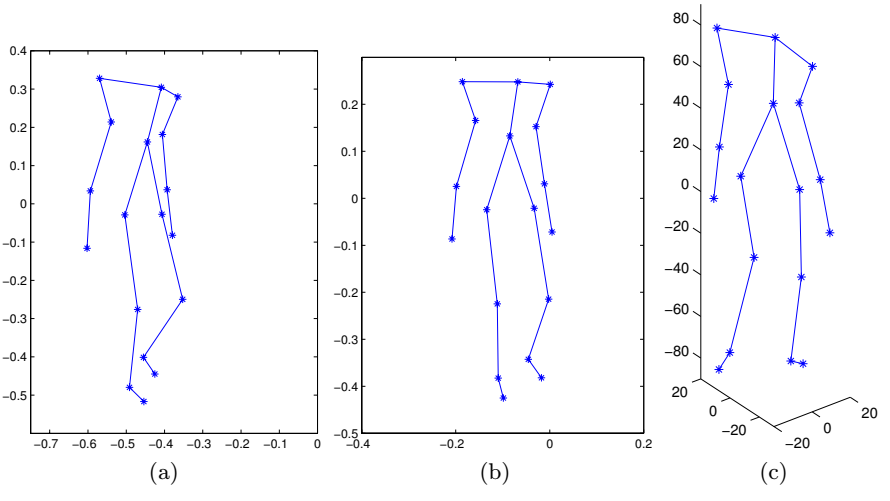


Fig. 2. Illustration of structure from periodic motion using motion capture data: (a) view at time t, (b) view at time $t + T_o$, (c) 3D reconstruction from T_o-separated views

3.3 Solving for Correspondences

In real video sequences, for which identified features are not available as in the mocap data, we can appeal to methods of interest point detection and corre-

spondence recovery that are used in conventional SFM. In particular, we use a RANSAC-based approach [17] on interest points extracted using the Förstner operator [3]. We perform interest point description and matching using the method of [18], which uses the L_1-norm on the error between vectors of filter responses computed at each interest point.

In using RANSAC to estimate the epipolar geometry, we assume that the feature points on the moving object dominate those in the rest of the scene. Because of this simplification, we do not need a separate figure/ground motion segmentation step as preprocessing.

3.4 Computing Dense Disparity

Once the epipolar geometry is known for an image pair, a number of dense stereo correspondence algorithms can be applied along the epipolar lines. In this work we use the method of [10], which is an energy minimization based method using a graph cut approximation. The input to the algorithm is a pair of rectified images (with respect to the object of interest) and the output is a disparity array. For rectification, we use the algorithm described in [8], Sec. 10.12.

4 Experiments

4.1 Walking Person I

Figure 3(c) shows the sparse 3D structure recovered for the T_o-separated frames of a walking person shown in Figure 3(a,b). A detail of the head and left shoulder region is shown in Figure 3(d) from a viewpoint behind the person and slightly to the left. Here we can see that the qualitative shape of the head relative to the sleeve region is reasonable.

The set of points used here consists of (i) the Förstner interest points used to estimate the fundamental matrix and (ii) the neighboring Canny edges with correspondences consistent with the epipolar geometry. Many points appear around the creases in the clothing, but this leaves several blank patches around the lower shirt and the arm.

4.2 Walking Person II

In Figure 4 we show an example of dense disparity estimation for another T_o-separated frame pair of a walking person. The input frames are shown at the top, followed by the rectified image pair. The estimated disparity relative to the left rectified image is shown next; for purposes of visualization, in this figure we have manually masked out the region corresponding to the person. The disparities are shown as a gray level, with lighter shades indicating larger disparity. We observe that the individual's right leg has higher disparity than the left leg, which is consistent with their depth ordering relative to the image plane, and that the majority of the disparity estimates for the rest of the body fall somewhere in between these values. In the original image pair, the light colored top of the forearm bleeds into the bright background; this corrupts the disparity estimate in that region.

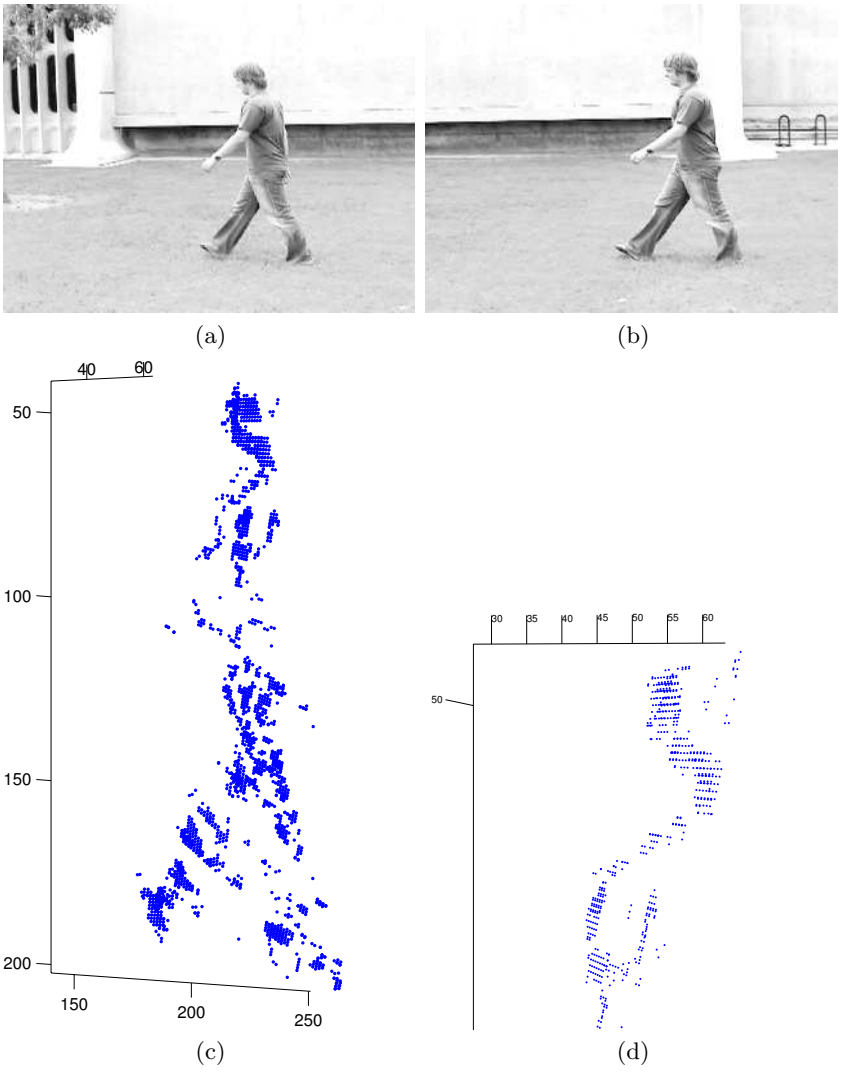


Fig. 3. (a,b) T_o-separated input frames. (c) Estimated 3D structure for interest points. (d) Detailed view of head and shoulder region viewed from behind the person.

4.3 Sensitivity Study

To conclude our experiments, we examine the sensitivity of the 3D reconstruction with respect to errors in the estimate of T_o. For this purpose, we again make use of the mocap data from Section 3.

We consider 200 frames of a regular walking sequence captured at 60 fps with $T_o \approx 90$ frames [16]. Each frame is a 2D projection (cf. Figure 2(a,b)) of the recorded 3D positions (which are accurate to 1mm) of a set of markers rigidly

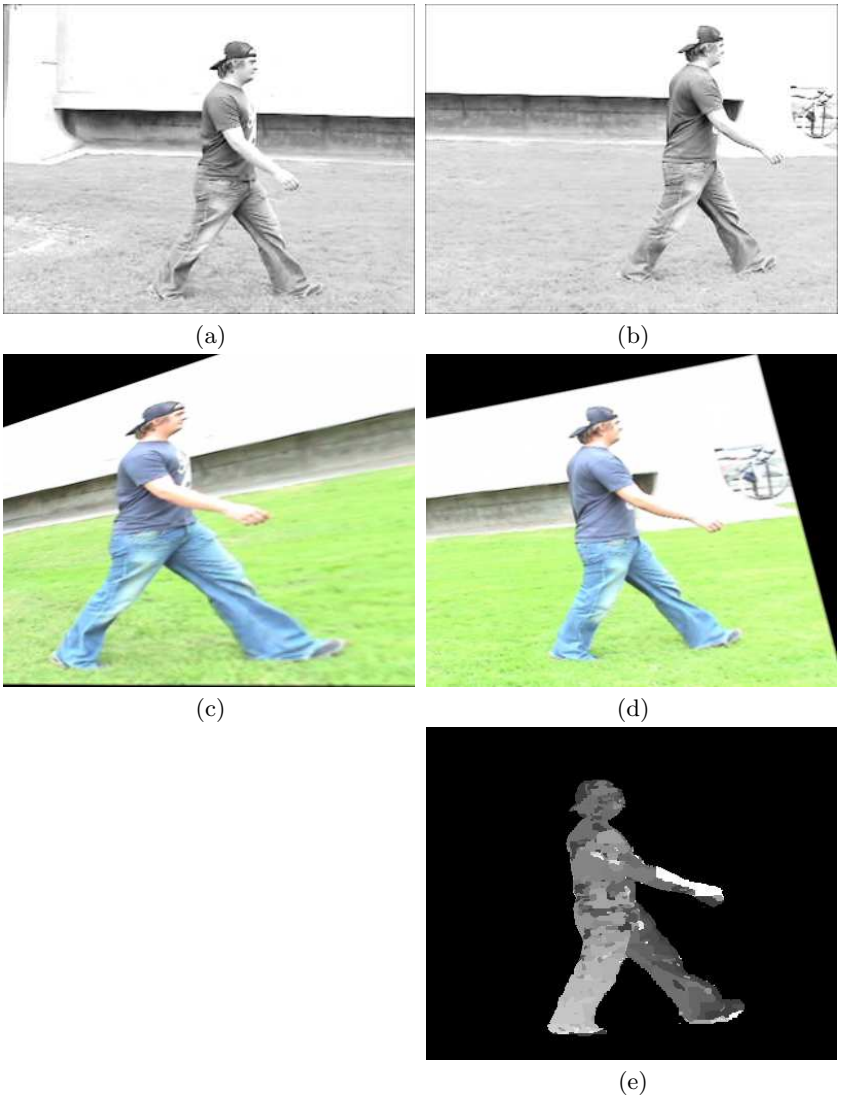
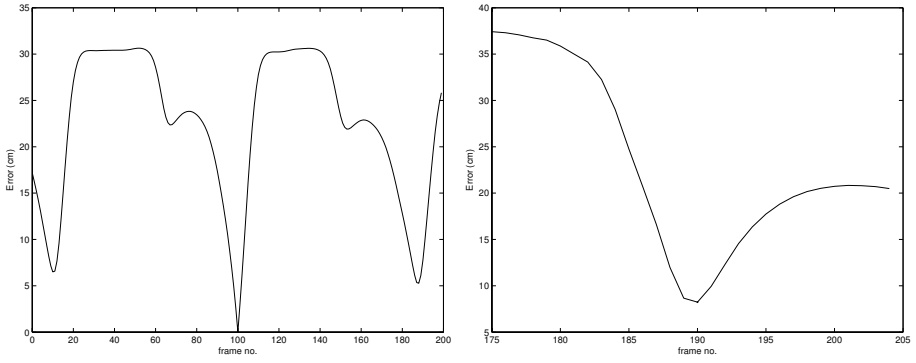


Fig. 4. (a,b) T_o-separated input frames. (c,d) Rectified images computed with respect to estimated epipolar geometry of input frames. (e) Estimated disparity, masked out to show region of interest containing the person.

attached to a subject's body. We selected a different 2D projection of frame 100 as a reference view. Using the reference view together with each of the previously mentioned 200 views, we computed the 3D reconstruction and the root-mean-square (RMS) error relative to the known 3D structure at the reference frame.

The error, which is plotted in Figure 5(a), is computed after solving for the least-squares homography aligning the projective reconstruction with the ground



(a) (b)

Fig. 5. Reconstruction error vs. frame number for mocap data of a walking person with $T_o \approx 90$ frames. (a) RMS error in units of cm between true 3D coordinates at frame 100 and the estimated 3D coordinates using one 2D view at frame 100 and a different 2D view at each of frames 1-200. (b) RMS error for frames 175-205 relative to frame 100, this time using the same 2D view for the reference frame as for frames 1-200.

truth marker positions at the reference frame. The periodicity is evident in the dips that occur at ± 90 frames on either side of 100. As expected, the error drops to zero at frame 100, at which point the reconstruction problem reduces to the case of exact stereo. The plot in Figure 5(b) shows a detail of the reconstruction error computed for 30 frames centered around frame 190; again the reference view is frame 100, but here the cameras specifying the 2D projections are the same for all the views. In each plot, it is evident that the error grows gradually with respect to displacements around the local optimum.

5 Conclusion and Future Work

We have presented an approach to 3D structure estimation based on monocular views of periodic motion. We demonstrated this approach using motion capture data and raw footage of pedestrians. Using the motion capture data, we explored the behavior of the reconstruction with respect to errors in the period estimation step.

The weakest part of the system is currently the correspondence estimation step. In theory, by the definition of periodicity, the problem treated in this work is identical to the classical SFM problem, provided the period estimate is correct. However, in practice, the correspondence problem is at least as hard as the usual stereo correspondence problem, and is in general harder, due to appearance variations across periods. In this regard, the correspondence problem associated with the SFPM problem lies somewhere in between the classical correspondence problem of wide-baseline stereo and the feature correspondence problem in 3D object recognition. We could therefore benefit from the use of methods designed

with the latter problem in mind; in future work we will investigate the use of scale-invariant keypoints [12] and affine invariant interest points [15].

Acknowledgments

We would like to thank Sameer Agarwal, Ben Ochoa, and Yi Ma for helpful discussions. We would also like to thank Claudio Fanti and Pietro Perona for providing the human motion capture data. This work was partially supported under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48 and by an NSF IGERT Grant (Vision and Learning in Humans and Machines, #DGE-0333451).

References

1. R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.
2. O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision, LNCS 588, Santa Margherita Ligure*, pages 563–578. Springer–Verlag, 1992.
3. W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ISPRS Intercommission Workshop*, Interlaken, June 1987.
4. A. R.J. François, G. G. Medioni, and R. Waupotitsch. Mirror symmetry $\Longrightarrow$ 2-view stereo geometry. *Image and Vision Computing*, 21(2):137–143, February 2003.
5. J. Gårding. Surface orientation and curvature from differential texture distortion. In *Proc. 5th Int'l Conf. on Computer Vision, Boston*, pages 733–739, 1995.
6. A.D. Gross and T. E. Boult. Analyzing skewed symmetries. *Int'l. Journal of Computer Vision*, 13(1):91–111, 1994.
7. R. I. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 1992.
8. Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
9. T. Kanade. Recovery of the three-dimensional shape of an object from a single view. *Artificial Intelligence*, 17:409–460, 1981.
10. Vladimir Kolmogorov and Ramin Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proc. 6th Int'l Conf. on Computer Vision, Vancouver*, 2001.
11. Y. Liu, R.T. Collins, and Y. Tsin. Gait sequence analysis using Frieze patterns. In *Proc. 7th Europ. Conf. Comput. Vision*, 2002.
12. D.G. Lowe. Demo software: Invariant keypoint detector. http://www.cs.ubc.ca/spider/lowe/keypoints/.
13. R. Koch M. Pollefeys and L. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. *Int'l. Journal of Computer Vision*, 32(1):7–25, 1999.
14. J. Malik and R. Rosenholtz. Computing local surface orientation and shape from texture for curved surfaces. *Int'l. Journal of Computer Vision*, 23(2):149–168, 1997.

15. Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142. Springer, 2002.
16. Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(7):814–827, 2003.
17. P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *Int Journal of Computer Vision*, 24(3):271–300, 1997.
18. Josh Wills, Sameer Agarwal, and Serge Belongie. What went where. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, volume 1, pages 37–44, 2003.